

UNITED STATES PATENT AND TRADEMARK OFFICE
DOCUMENT CLASSIFICATION BARCODE SHEET



CATEGORY:

CLEARED

ADDRESS
CONTACT IF FOUND:



07/31/00

Please type a plus sign (+) inside this box → Approved for use through 09/30/00. OMB 0651-0032
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No. K35A0648 Total Pages 23

First Named Inventor or Application Identifier

CHRISTOPHER L. HAMLIN

Express Mail Label No. EI266879775US

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. Fee Transmittal Form
(Submit an original, and a duplicate for fee processing)
2. Specification [Total Pages 16]
(preferred arrangement set forth below)
 - Descriptive title of the Invention
 - Cross References to Related Applications
 - Statement Regarding Fed sponsored R & D
 - Reference to Microfiche Appendix
 - Background of the Invention
 - Brief Summary of the Invention
 - Brief Description of the Drawings (if filed)
 - Detailed Description
 - Claim(s)
 - Abstract of the Disclosure
3. Drawing(s) (35 USC 113) [Total Sheets 5]
 X Formal Informal
4. Oath or Declaration [Total Pages]
 - a. Newly executed (original or copy)
 - b. Copy from a prior application (37 CFR 1.63(d))
(for continuation/divisional/with Box 17 completed)
[Note Box 5 below]
 - i. DELETION OF INVENTOR(S)
Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. Incorporation By Reference (useable if Box 4b is checked)
The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.

ADDRESS TO: Assistant Commissioner for Patents
Box Patent Application
Washington, DC 20231

6. Microfiche Computer Program (Appendix)
7. Nucleotide and/or Amino Acid Sequence Submission
(if applicable, all necessary)
 - a. Computer Readable Copy
 - b. Paper Copy (identical to computer copy)
 - c. Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. Assignment Papers (cover sheet & document(s))
9. 37 CFR 3.73(b) Statement Power of Attorney
(when there is an assignee)
10. English Translation Document (if applicable)
11. Information Disclosure Statement (IDS)/PTO-1449 Copies of IDS Citations
12. Preliminary Amendment
13. Return Receipt Postcard (MPEP 503)
(Should be specifically itemized)
14. Small Entity Statement filed in prior application, Statement(s) Status still proper and desired
15. Certified Copy of Priority Document(s)
(if foreign priority is claimed)
16. Other:

17. If a CONTINUING APPLICATION, check appropriate box and supply the requisite information:
 Continuation Divisional Continuation-in-part (CIP) of prior application No: _____ / _____
18. CORRESPONDENCE ADDRESS Customer Number or Bar Code Label

(Insert Customer No. or Attach bar code label here)

or Correspondence address below

NAME	WESTERN DIGITAL CORPORATION		
	Milad G. Shara, Esq. - Reg. 39,367		
ADDRESS	8105 IRVINE CENTER DRIVE		
	PLAZA 3		
CITY	IRVINE	STATE	CALIFORNIA ZIP CODE 92618
COUNTRY	U.S.A.	TELEPHONE	(949) 932-5676 FAX (949) 932-5633

Burden Hour Statement: This form is estimated to take 0.2 hours to complete. Time will vary depending upon the needs of the individual case. Any comments on the amount of time you are required to complete this form should be sent to the Chief Information Officer, Patent and Trademark Office, Washington, DC 20231. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Assistant Commissioner for Patents, Box Patent Application, Washington, DC 20231.

1 **TRANSFERRING SCHEDULING DATA FROM A PLURALITY OF DISK STORAGE**
2 **DEVICES TO A NETWORK SWITCH BEFORE TRANSFERRING DATA**
3 **ASSOCIATED WITH SCHEDULED REQUESTS BETWEEN THE NETWORK**
4 **SWITCH AND A PLURALITY OF HOST INITIATORS**

5

6 **BACKGROUND OF THE INVENTION**

7

8 **Field of the Invention**

9 The present invention relates to network systems. More particularly, the present
10 invention relates to transferring scheduling data from a plurality of disk storage devices to a
11 network switch before transferring data associated with scheduled requests between the network
12 switch and a plurality of host initiators.

13 **Description of the Prior Art**

14 Conventional disk drive storage systems typically employ a scheduling algorithm in order
15 to optimize data throughput. For example, a scheduling algorithm may evaluate and prioritize
16 access requests rather than service the requests on a "first come first serve" basis. The
17 scheduling priority is typically based on certain temporal parameters of the disk drive, such as
18 the radial position of the head with respect to the disk. A scheduling algorithm may, for
19 example, service all of the access requests from the outer to inner diameter tracks before
20 servicing access requests from the inner to outer diameter tracks, similar to an elevator in a
21 building servicing all of the down requests before servicing up requests. This algorithm is
22 appropriately referred to as the "elevator" algorithm.

23 It is known to use temporal parameters of a disk drive (e.g., the radial or circumferential
24 position of the head) in order to perform the scheduling operations; however, these types of
25 scheduling algorithms have in the past been implemented by a disk controller which has direct
26 access to the temporal parameters. For example, U.S. Patent No. 5,390,313 discloses a disk drive
27 comprising a disk controller for scheduling access to multiple disks based on the circumferential

1 position of the heads relative to the disks. Co-pending U.S. Patent Application Serial No.
2 09/301,179 discloses a disk drive which provides head position information to a host computer
3 so that the host microprocessor may execute the scheduling algorithm rather than the disk
4 controller. U.S. Patent No. 5,787,482 discloses a video server wherein access requests to a
5 plurality of disk drives are scheduled based on an inferred radial position of the head within each
6 disk drive. The radial position of the head is inferred based on commands previously sent to
7 each disk drive. However, using inferred temporal parameters to implement the scheduling
8 algorithm provides sub-optimal performance due to the error inherent in estimation. Further, it is
9 difficult to minimize the variance in latency associated with generating the temporal parameters
10 due to the estimation error as well as the variance in computing the temporal parameters, which
11 further degrades performance of the scheduling algorithm. Consequently, scheduling algorithms
12 based on inferred temporal parameters are sub-optimal with respect to the aggregate performance
13 of a computer network, and particularly the number of input/output operations per second (IOPs)
14 performed by each disk drive connected to the computer network.

15 There is, therefore, a need to improve upon the prior art techniques of scheduling access
16 to a plurality of storage systems, such as a plurality of disk storage devices, connected to a
17 computer network. In particular, there is a need to schedule access to a plurality of disk storage
18 devices connected to a computer network in a manner which minimizes the variance in latency
19 associated with generating the temporal parameters, thereby improving the computer network's
20 aggregate performance.

21 **SUMMARY OF THE INVENTION**

22 The present invention may be regarded as a network switch for resolving requests from a
23 plurality of host initiators by scheduling access to a plurality of disk storage devices. The
24 network switch comprises a switched fabric comprising a plurality of switching elements. Each
25 switching element comprises a plurality of bi-directional switched fabric ports, and a control
26 input connected to receive switch control data for selectively configuring the switching element
27 in order to interconnect the bi-directional switched fabric ports. The network switch further

1 comprises a memory for storing a routing and scheduling program, and a microprocessor,
2 responsive to the requests, for executing the steps of the routing and scheduling program to
3 generate the switch control data to transmit scheduled requests through the bi-directional
4 switched fabric ports. At least one of the plurality of switching elements comprises a disk
5 storage interface for connecting to a selected one of the disk storage devices. The
6 microprocessor schedules access to the plurality of disk storage devices through the disk storage
7 interface. The disk storage interface receives scheduling data from the selected one of the
8 storage devices, and the memory stores the scheduling data received via the bi-directional
9 switched fabric ports of a selected number of the switching elements. The scheduling data is
10 processed according to a priority such that the selected switching elements transfer the
11 scheduling data through the bi-directional switched fabric ports before transferring data
12 associated with the scheduled requests.

13 In one embodiment, the at least one switching element further comprise a disk storage
14 device connected to the disk storage interface. In another embodiment, the switching elements
15 further comprise a plurality of virtual lanes, wherein at least one of the virtual lanes is reserved
16 for transferring data associated with the scheduled requests, at least one of the virtual lanes is
17 reserved for transferring the scheduling data, and the virtual lane for transferring the scheduling
18 data comprises a higher priority than the virtual lane for transferring the data associated with the
19 scheduled requests.

20 The present invention may also be regarded as a method of resolving requests from a
21 plurality of host initiators by scheduling access to a plurality of disk storage devices connected to
22 a network switch, the network switch comprising a switched fabric comprising a plurality of
23 switching elements. The method comprises the steps of transmitting through the switching
24 elements scheduling data from the plurality of disk storage devices to a memory, evaluating the
25 scheduling data in order to schedule the requests from the host initiators, and transmitting data
26 associated with the scheduled requests through the switching elements to the plurality of disk
27 storage devices. The scheduling data is processed according to a priority such that the switching

1 elements transfer the scheduling data before transferring data associated with the scheduled
2 requests.

3 **BRIEF DESCRIPTION OF THE DRAWINGS**

4 FIG. 1 shows a network switch according to an embodiment of the present invention
5 comprising a switched fabric having a plurality of switching elements, a microprocessor for
6 executing a routing and scheduling algorithm, and a memory for storing scheduling data received
7 from a plurality of disk storage devices.

8 FIG. 2 shows an example topology for the switched fabric of FIG. 1 wherein the plurality
9 of multi-port switches form a multi-dimensional switched fabric.

10 FIG. 3 shows an embodiment of the present invention wherein each switching element in
11 the switched fabric of FIG. 2 comprises a disk storage device (DSD) to form a switch storage
12 node.

13 FIG. 4 illustrates details of a switch storage node for use in the embodiment of FIG. 3.

14 FIG. 5 shows details of the disk storage device (DSD) and disk storage interface circuitry
15 of FIG. 4.

16 FIG. 6 shows an alternative embodiment of the present invention wherein the disk storage
17 device (DSD) is implemented external to the switching element as a disk drive connected to the
18 switch fabric.

19 **DESCRIPTION OF THE PREFERRED EMBODIMENTS**

20 FIG. 1 shows a network switch 2 for resolving requests from a plurality of host initiators
21 by scheduling access to a plurality of disk storage devices (DSDs). The network switch 2
22 comprises a switched fabric 4 comprising a plurality of switching elements. Each switching
23 element comprises a plurality of bi-directional switched fabric ports, and a control input
24 connected to receive switch control data for selectively configuring the switching element in
25 order to interconnect the bi-directional switched fabric ports. The network switch 2 further
26 comprises a memory 6 for storing a routing and scheduling program, and a microprocessor 8,
27 responsive to the requests, for executing the steps of the routing and scheduling program to

1 generate the switch control data to transmit scheduled requests through the bi-directional
2 switched fabric ports. At least one of the plurality of switching elements comprises a disk
3 storage interface for connecting to a selected one of the disk storage devices. The
4 microprocessor 8 schedules access to the plurality of disk storage devices through the disk
5 storage interface. The disk storage interface receives scheduling data from the selected one of
6 the storage devices, and the memory 6 stores the scheduling data received via the bi-directional
7 switched fabric ports of a selected number of the switching elements. The scheduling data is
8 processed according to a priority such that the selected switching elements transfer the
9 scheduling data through the bi-directional switched fabric ports before transferring data
10 associated with the scheduled requests.

11 The microprocessor 8 in the network switch 2 executes a conventional routing algorithm
12 for routing requests (messages) between the nodes in the network (e.g., host initiators, storage
13 devices, etc.). The network switch 2 comprises buffers 10₀-10_N which buffer the requests before
14 and after the requests are transmitted through the switched fabric 4. In one embodiment, a
15 request consists of a packet having a packet header comprising routing data which identifies the
16 source node for the packet. The microprocessor 8 processes the packet header in order to route
17 the packet through the switched fabric 4. A suitable routing algorithm implemented by the
18 microprocessor 8 generates control data for configuring the switching elements within the
19 switched fabric 4. Any suitable routing algorithm may be implemented by the network switch 2,
20 and it may support Unicast or Multicast Routing. The routing decisions may be made centrally,
21 at the source, distributed, or multiphase, implemented using a lookup table or using a finite-state
22 machine. Further, the routing algorithm may be deterministic or adaptive. A discussion of
23 various routing algorithms which may be employed in the embodiments of the present invention
24 is provided by Jose Duato et al. in the text book "*Interconnection Networks, an Engineering*
25 *Approach*", IEEE Computer Society, 1997.

26 The routing algorithm is implemented a layer "above" the switching layer, and thus the
27 routing algorithm may be compatible with various different switching algorithms, for example,

1 Virtual Cut-Through Switching, Wormhole Switching, and Mad Postman Switching. The
2 switching layer is implemented by the switched fabric 4 using a plurality of multi-port switching
3 elements. FIG. 2 illustrates an example topology for the switching elements: a two dimensional
4 switched fabric which allows any node in the network to communicate with any other node so
5 that many nodes can communicate simultaneously without contention. In an alternative
6 embodiment, the switching elements are configured to form a plurality of switch stages, wherein
7 each individual switch stage is a multi-dimensional switched fabric, and the number of switch
8 stages and connection patterns between switch stages determines the routing capability of the
9 network switch 2. In the two dimensional switched fabric of FIG. 2, each switching element
10 comprises up to four ports (North, South, East and West); however, switching elements
11 comprising fewer or more ports may also be employed. Various topologies and switching
12 algorithms which may be employed in the embodiments of the present invention are discussed in
13 the aforementioned text book by Jose Duato et al..

14 FIG. 3 shows an embodiment of the present invention wherein each switching element in
15 the switched fabric 4 of FIG. 1 is implemented as a switch storage node 12 comprising a local
16 disk storage device (DSD) 14. Each DSD 14 comprises at least one disk for storing data, and a
17 corresponding head actuated radially over the disk for writing data to and reading data from the
18 disk. In the embodiment of FIG. 3, each storage node comprises four bi-directional ports
19 (N,E,S,W) which interconnect with other switch storage nodes 12 to form the two dimensional
20 switched fabric of FIG. 2. The dimensions in the switched fabric can be increased by increasing
21 the number of ports at each storage node. For example, a three dimensional switched fabric
22 comprises storage nodes having six bi-directional ports. Control data 16 generated by the
23 microprocessor 8 of FIG. 1 configures each switch storage node 12 in order to implement the
24 routing and scheduling algorithm. In one embodiment, the switched fabric of FIG. 3 implements
25 a distributed file system wherein file data is mirrored on multiple DSDs 14 to enhance
26 performance. The host initiators of FIG. 1 request access to specific files, and the
27 microprocessor 8 evaluates the requests in view of the current state of the switched fabric 4 to

1 determine the most appropriate path and DSD 14 to service the requests. To this end, the DSDs
2 14 in each switch storage node 12 of FIG. 3 periodically transfer scheduling data autonomously
3 to the memory 6 of FIG. 1 for use by the microprocessor 8 in routing and scheduling the requests
4 received from the host initiators.

5 FIG. 4 shows details of a switch storage node 12 according to one embodiment of the
6 present invention for use in the switched fabric 4 of FIG. 3. Each port (N,E,S,W) in FIG. 4
7 comprises an input port (18A-18D) and an output port (20A-20D) which facilitates the bi-
8 directional aspect of each port. Data is received into the switch storage node 12 via the input
9 ports (18A-18D), stored in a data buffer 22, and then routed to a selected output port or ports
10 (20A-20D) at the appropriate time. Header information is extracted from the input data and input
11 into a routing table 24 which comprises the routing information as configured by the control data
12 16. A scheduler 26 processes the selections 28 made by the routing table 24 in order to transfer
13 the data from the data buffer 22 to the appropriate output port (20A-20D) at the appropriate time.
14 In one embodiment, the switch storage node 12 may support isochronous data wherein the data
15 stored in the data buffer 22 is transferred to the appropriate output port (20A-20D) according to
16 an arrival and deadline time which guarantees a maximum delay time for the data to cross the
17 switch storage node 12.

18 Each output port (20A-20D) comprises a plurality of virtual lanes or queues (e.g., 22A
19 and 22B) which are prioritized so that the data stored in the virtual lanes having higher priority
20 are transferred over data stored in virtual lanes having lower priority. The prioritized virtual
21 lanes are used to transmit scheduling data associated with the DSD 14 prior to sending data
22 associated with host initiator requests in order to minimize the latency in transmitting the
23 scheduling data to the memory 6 of FIG. 1.

24 The switched storage node 12 of FIG. 4 comprises disk storage interface circuitry 23 for
25 interfacing with the DSD 14. Data received from the input ports (18A-18D) and destined for
26 storage on the DSD 14 is transmitted via the scheduler 26 through the disk storage interface 23
27 and written on a disk within the DSD 14. A request to read data stored on the DSD 14 may also

1 be received from the input ports (18A-18D). The request is transferred by the scheduler 26 to the
2 disk storage interface 23 which interfaces with the DSD 14 over line 25 to perform the read
3 operation. The data read from the DSD 14 is configured by the disk storage interface 23 into
4 network data (e.g., network packets) which is injected into the switching circuitry similar to data
5 received from the input ports (18A-18D). The data is staged in the data buffer 22 and header
6 information is transferred to the routing table 24 for use in routing the data to the appropriate
7 output port (20A-20D).

8 To assist the microprocessor 8 of FIG. 1 with the routing and scheduling of requests
9 received from the host initiators, the DSD 14 in each of the switching nodes 12 periodically and
10 autonomously transfers scheduling data to the memory 6 of FIG. 1. In one embodiment, the
11 scheduling data comprises temporal parameters of the DSD 14 which provide insight into the
12 current state of the DSD 14 and latency associated with storing or retrieving particular data (e.g.,
13 a data stream). Examples of temporal data include the radial location of the head within the DSD
14 relative to the disk, as well as the circumferential position of the head relative to the disk.

15 FIG. 5 shows an embodiment of the present invention wherein the DSD 14 comprises
16 components of a conventional head disk assembly (HDA), including a head 28 positioned
17 radially over a disk 30 by a voice coil motor (VCM) 32 which rotates an actuator arm 34 about a
18 pivot. An index mark 36 is recorded on the disk 30 which provides information about the
19 circumferential location of the head 28 relative to the disk 30. The disk storage interface
20 circuitry 23 comprises a VCM driver 36 for generating control signals 38 applied to the VCM 32,
21 and a servo control system 40 for generating control signals 42 applied to the VCM driver 36. A
22 disk controller 44 within the disk storage interface 23 receives requests from the switching
23 circuitry over line 25 to write data to and read data from the disk 30. The disk controller 44
24 generates control signals 48 applied to the servo control system 40 in order to position the head
25 28 over a desired radial location of the disk 30.

26 During a write operation, the disk controller 44 receives a request over line 25 to write
27 data to the disk 30. The disk controller 44 evaluates the request to determine the appropriate

1 radial location on the disk 30 to write the data, and then positions the head 28 over the desired
2 radial location by sending the appropriate control signals 48 to the servo control system 40. The
3 write data is formatted (e.g., encoded using an error correction code (ECC), defect mapped, etc.),
4 and then transmitted over line 52 for writing to the disk 30. During a read operation, the disk
5 controller 44 processes the request received over line 25 by positioning the head 28 over the
6 desired radial location of the disk 30 and reading the data over line 52. The disk controller 44
7 configures the data read from the disk 30 into network data (e.g., network packets) which is
8 transmitted to the switching circuitry of FIG. 4 over line 25.

9 A register file 50 is employed in the embodiment of FIG. 5 for storing scheduling data in
10 the form of temporal parameters of the DSD 14. For example, in one embodiment the servo
11 control system 40 stores in the register file 50 the current radial and circumferential location of
12 the head 28 with respect to the disk 30. The radial location of the head 28 is determined from
13 Gray coded track addresses in embedded servo sectors recorded at a regular interval on the disk
14 30, and the circumferential location of the head 28 is determined relative to when the head 28
15 reads the index mark 36. At a predetermined periodic interval, the disk controller 44 retrieves
16 the scheduling data from the register file 50, converts the data into network data (e.g., a network
17 packet), and transmits the network data over line 25 to the switching circuitry for transfer to the
18 memory 6 of FIG. 1.

19 In order to minimize the latency associated with transferring the scheduling data to the
20 memory 6, the scheduling data is transmitted using a virtual lane (e.g., 22A or 22B of FIG. 4)
21 which has a higher priority than regularly scheduled data associated with requests from the host
22 initiators. This ensures the scheduling data will not be delayed in the network by regularly
23 scheduled host initiator data.

24 In another embodiment, the scheduling data is transferred according to an isochronous
25 protocol to ensure that the scheduling data arrives at the memory 6 within a specified period.
26 This minimizes the variance in the latency associated with transmitting the scheduling data to the
27 memory 6. An example of an isochronous protocol is provided in the reference "FireWire®

1 System Architecture, Second Edition IEEE 1394a", MindShare, Inc., 1999.

2 In yet another embodiment, a primary network communication protocol is used to
3 communicate with the host initiators in order to transmit host initiator data through the network,
4 and a secondary network communication protocol is used to communicate, at least in part, with
5 the disk storage devices. For example, the primary network communication protocol may
6 implement an asynchronous protocol for communicating host initiator data, and the secondary
7 network communication protocol may implement an isochronous protocol for communicating
8 the drive scheduling data. In another embodiment, different switching techniques may be
9 employed to implement the dual level protocol. For example, the primary network
10 communication protocol may implement wormhole switching in order to communicate long
11 packets associated with host initiator data more efficiently, whereas the secondary network
12 communication protocol may implement packet switching in order to communicate short packets
13 of scheduling data more efficiently. With packet switching, packets are buffered completely
14 within each node before being transferred to the next node which is why this type of switching is
15 referred to as store-and-forward packet switching. With wormhole switching, packets are
16 transmitted immediately to the next node so that packets are pipelined across several nodes.
17 Packet switching ensures a predictable consumption of link and buffer resources which is
18 necessary to support an isochronous communication protocol, whereas wormhole switching
19 reduces the latency and buffer requirements needed to transfer the typically longer packets
20 associated with host initiator data.

21 In the embodiment shown in FIG. 4, the disk interface circuitry of FIG. 5 and switching
22 circuitry are integrated into a single integrated circuit (IC) which is mounted on a printed circuit
23 board (PCB) attached to a conventional HDA. In another embodiment illustrated in FIG. 6, the
24 switching circuitry and disk storage interface circuitry are implemented separate from the disk
25 controller 44 and other circuitry shown in FIG. 5. In this embodiment, the disk storage interface
26 circuitry may comprise the facilities for converting between network and DSD data, whereas the
27 disk controller 44 and other circuitry shown in FIG. 5 may be implemented within the DSD 14.

- 1 For example, the DSD 14 may be implemented as a conventional disk drive modified to collect
- 2 and transmit scheduling data at a periodic interval. The disk drive may be connected at a node
- 3 within the switched fabric 4, or it may be connected at the edge of the switched fabric 4 as
- 4 illustrated in FIG. 1.

1 **WE CLAIM:**

- 1 1. A network switch for resolving requests from a plurality of host initiators by scheduling
2 access to a plurality of disk storage devices, the network switch comprising:
3 (a) a switched fabric comprising a plurality of switching elements, each switching
4 element comprising:
5 a plurality of bi-directional switched fabric ports; and
6 a control input connected to receive switch control data for selectively configuring
7 the switching element in order to interconnect the bi-directional switched
8 fabric ports;
9 (b) a memory for storing a routing and scheduling program; and
10 (c) a microprocessor, responsive to the requests, for executing the steps of the routing
11 and scheduling program to generate the switch control data to transmit scheduled
12 requests through the bi-directional switched fabric ports,
13 wherein:
14 at least one of the plurality of switching elements comprises a disk storage interface
15 for connecting to a selected one of the disk storage devices;
16 the microprocessor for scheduling access to the plurality of disk storage devices
17 through the disk storage interface;
18 the disk storage interface for receiving scheduling data from the selected one of the
19 storage devices;
20 the memory for receiving the scheduling data via the bi-directional switched fabric
21 ports of a selected number of the switching elements; and
22 the scheduling data is processed according to a priority such that the selected
23 switching elements transfer the scheduling data through the bi-directional
24 switched fabric ports before transferring data associated with the scheduled
25 requests.

- 1 2. The network switch as recited in claim 1, wherein the at least one switching element
2 further comprises a disk storage device connected to the disk storage interface.
 - 1 3. The network switch as recited in claim 1, wherein:
 - 2 (a) each disk storage device comprises a disk and a head; and
 - 3 (b) the scheduling data comprises a radial location of the head relative to the disk within
4 each disk storage device.
 - 1 4. The network switch as recited in claim 3, wherein the scheduling data further comprises a
2 circumferential location of the head relative to the disk within each disk drive.
 - 1 5. The network switch as recited in claim 1, wherein the switching elements further
2 comprise a plurality of virtual lanes, wherein:
 - 3 (a) at least one of the virtual lanes is reserved for transferring data associated with the
4 scheduled requests;
 - 5 (b) at least one of the virtual lanes is reserved for transferring the scheduling data; and
 - 6 (c) the virtual lane for transferring the scheduling data comprises a higher priority than
7 the virtual lane for transferring the data associated with the scheduled requests.
 - 1 6. The network switch as recited in claim 1, wherein the scheduling data is communicated to
2 the memory through the bi-directional switched fabric ports according to an isochronous
3 protocol.

1 7. A method of resolving requests from a plurality of host initiators by scheduling access to
2 a plurality of disk storage devices connected to a network switch, the network switch
3 comprising a switched fabric comprising a plurality of switching elements, the method
4 comprising the steps of:

- 5 (a) transmitting through the switching elements scheduling data from the plurality of
6 disk storage devices to a memory;
7 (b) evaluating the scheduling data in order to schedule the requests from the host
8 initiators; and
9 (c) transmitting data associated with the scheduled requests through the switching
10 elements to the plurality of disk storage devices,

11 wherein the scheduling data is processed according to a priority such that the
12 switching elements transfer the scheduling data before transferring data associated
13 with the scheduled requests.

1 8. The method as recited in claim 7, wherein:

- 2 (a) each disk storage device comprises a disk and a head; and
3 (b) the scheduling data comprises a radial location of the head relative to the disk within
4 each disk storage device.

1 9. The method as recited in claim 8, wherein the scheduling data further comprises a
2 circumferential location of the head relative to the disk within each disk drive.

1 10. The method as recited in claim 7, wherein the switching elements further comprise a
2 plurality of virtual lanes, wherein:

- 3 (a) at least one of the virtual lanes is reserved for transferring data associated with the
4 scheduled requests;
5 (b) at least one of the virtual lanes is reserved for transferring the scheduling data; and

6 (c) the virtual lane for transferring the scheduling data comprises a higher priority than
7 the virtual lane for transferring the data associated with the scheduled requests.

1 11. The method as recited in claim 7, wherein the scheduling data is communicated to the
2 memory through the switching elements according to an isochronous protocol.

1 **TRANSFERRING SCHEDULING DATA FROM A PLURALITY OF DISK STORAGE**
2 **DEVICES TO A NETWORK SWITCH BEFORE TRANSFERRING DATA**
3 **ASSOCIATED WITH SCHEDULED REQUESTS BETWEEN THE NETWORK**
4 **SWITCH AND A PLURALITY OF HOST INITIATORS**

5

6 **ABSTRACT OF THE DISCLOSURE**

7 A network switch is disclosed for resolving requests from a plurality of host initiators by
8 scheduling access to a plurality of disk storage devices. The network switch comprises a
9 switched fabric comprising a plurality of switching elements. Each switching element comprises
10 a plurality of bi-directional switched fabric ports, and a control input connected to receive switch
11 control data for selectively configuring the switching element in order to interconnect the bi-
12 directional switched fabric ports. The network switch further comprises a memory for storing a
13 routing and scheduling program, and a microprocessor, responsive to the requests, for executing
14 the steps of the routing and scheduling program to generate the switch control data to transmit
15 scheduled requests through the bi-directional switched fabric ports. At least one of the plurality
16 of switching elements comprises a disk storage interface for connecting to a selected one of the
17 disk storage devices. The microprocessor schedules access to the plurality of disk storage
18 devices through the disk storage interface. The disk storage interface receives scheduling data
19 from the selected one of the storage devices, and the memory stores the scheduling data received
20 via the bi-directional switched fabric ports of a selected number of the switching elements. The
21 scheduling data is processed according to a priority such that the selected switching elements
22 transfer the scheduling data through the bi-directional switched fabric ports before transferring
23 data associated with the scheduled requests.

FIG. 1

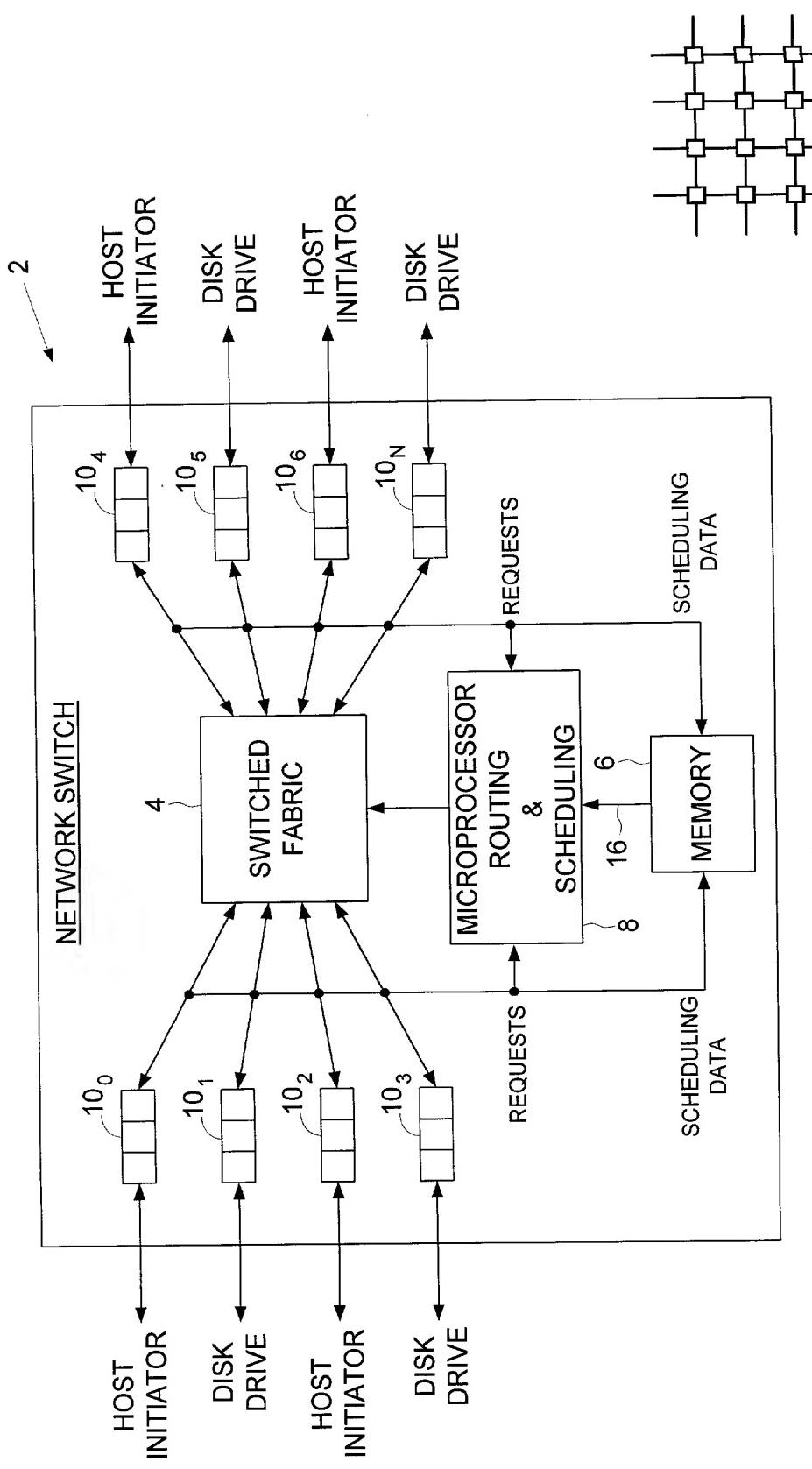
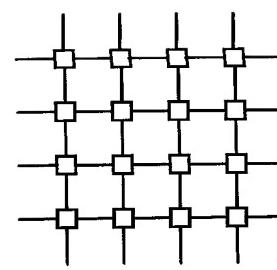


FIG. 2



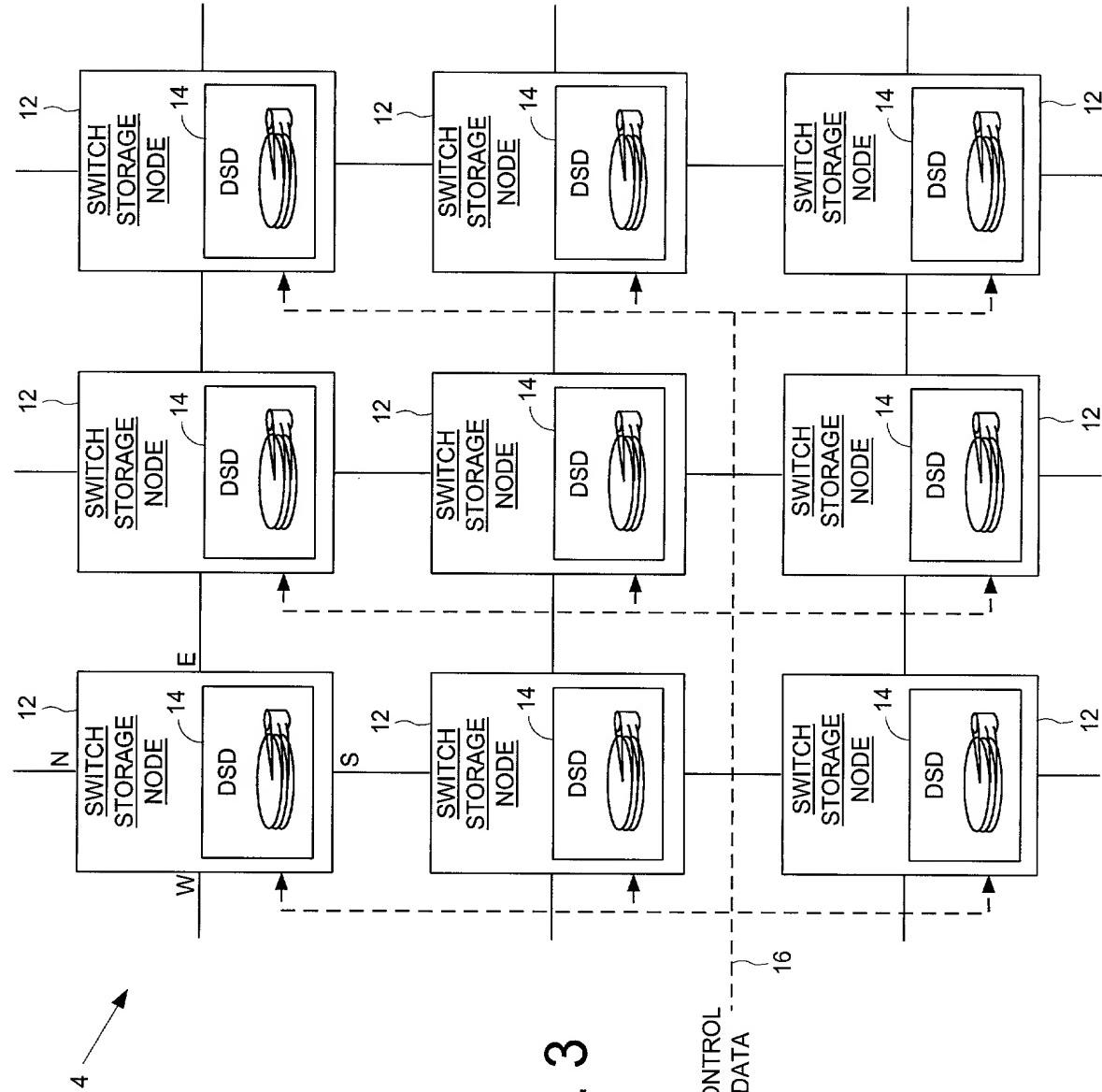


FIG. 3

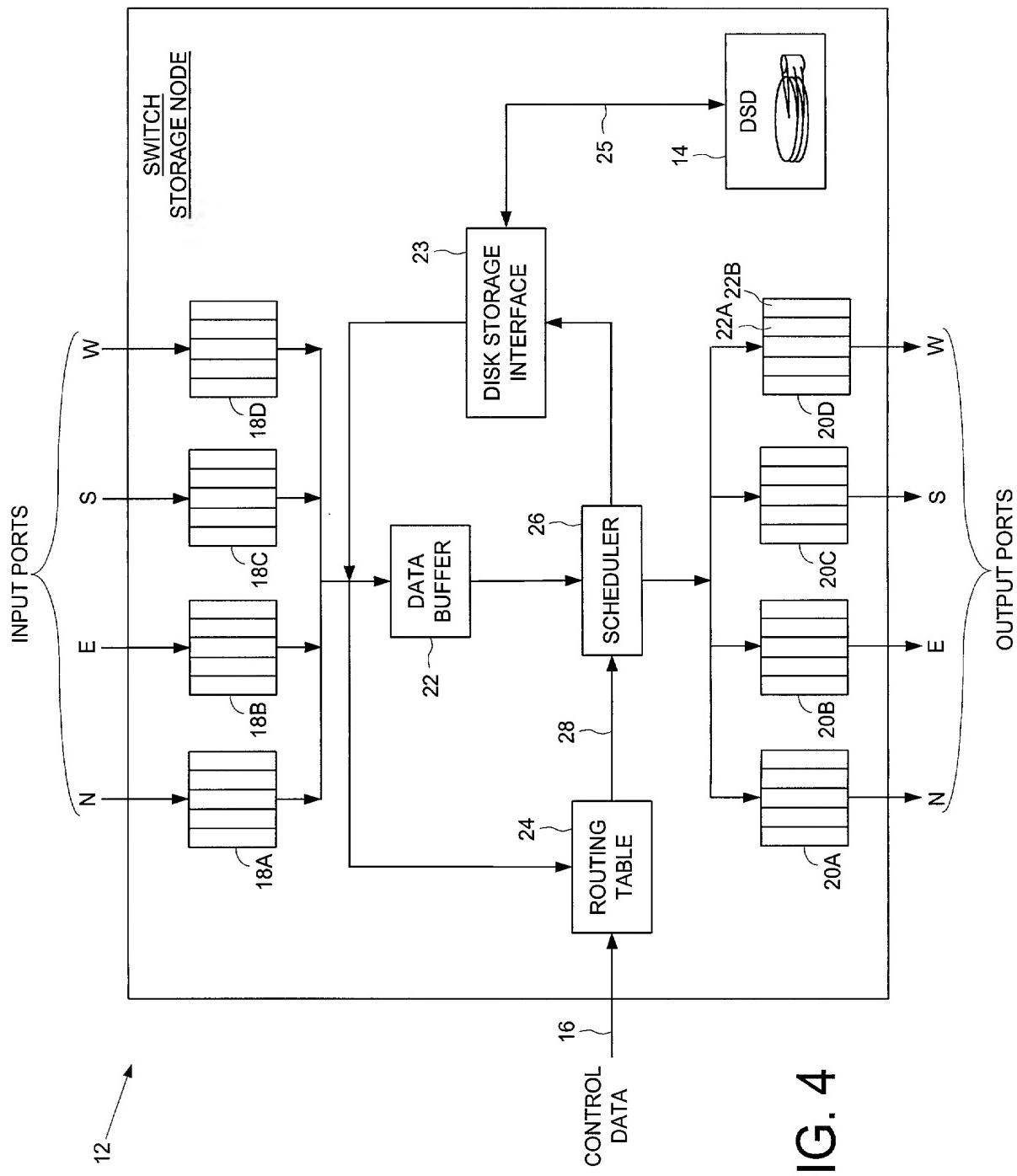
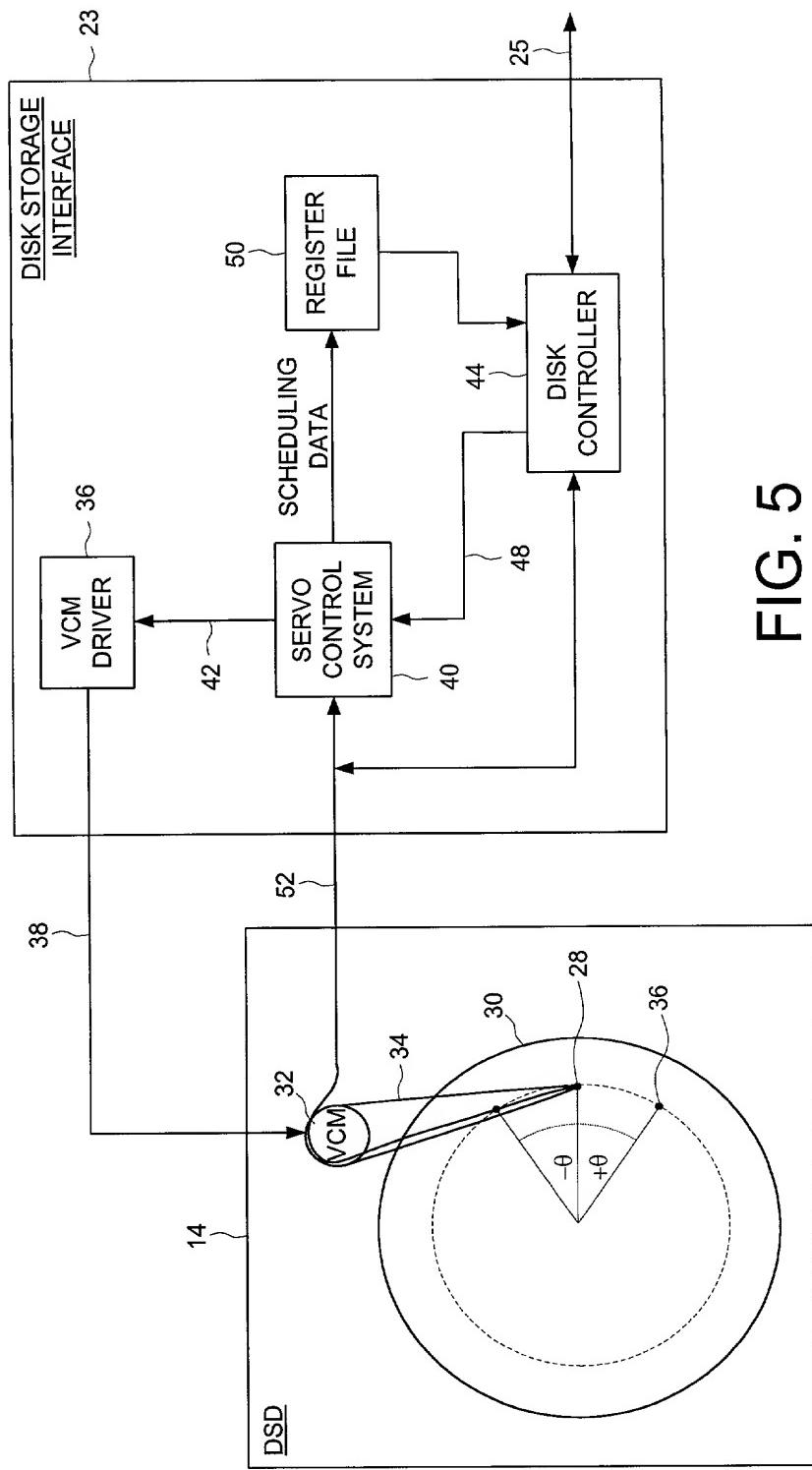


FIG. 4

FIG. 5



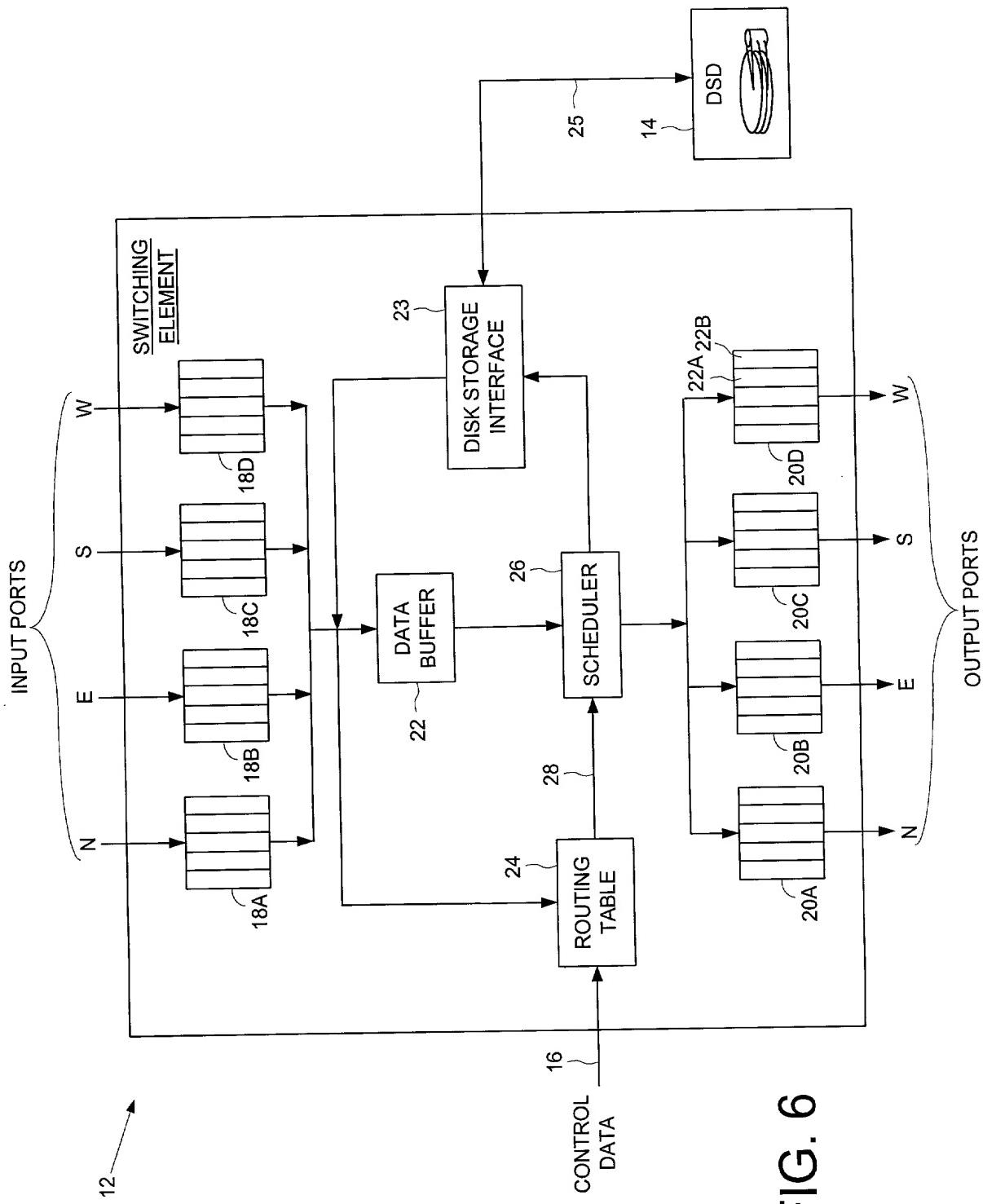


FIG. 6